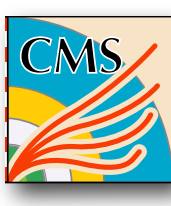


High Level View From Experiments

Michael Ernst, Ian Fisk

May 14, 2008

HEP Computing

Full Simulation

KHz
Generation
mHz
Simulation
Hz
Digitization

Fast Simulation

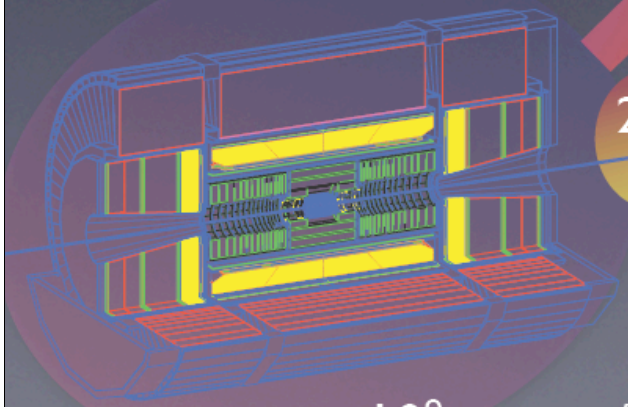
KHz
Generation
Hz
Fast Simulation

Data Store

Balance of full to fast sim varies

cHz
Reconstruction

High-level Trigger



200/300 Hz

10^9 events/year

Data Base

KHz
Hz
Algorithmic Analysis

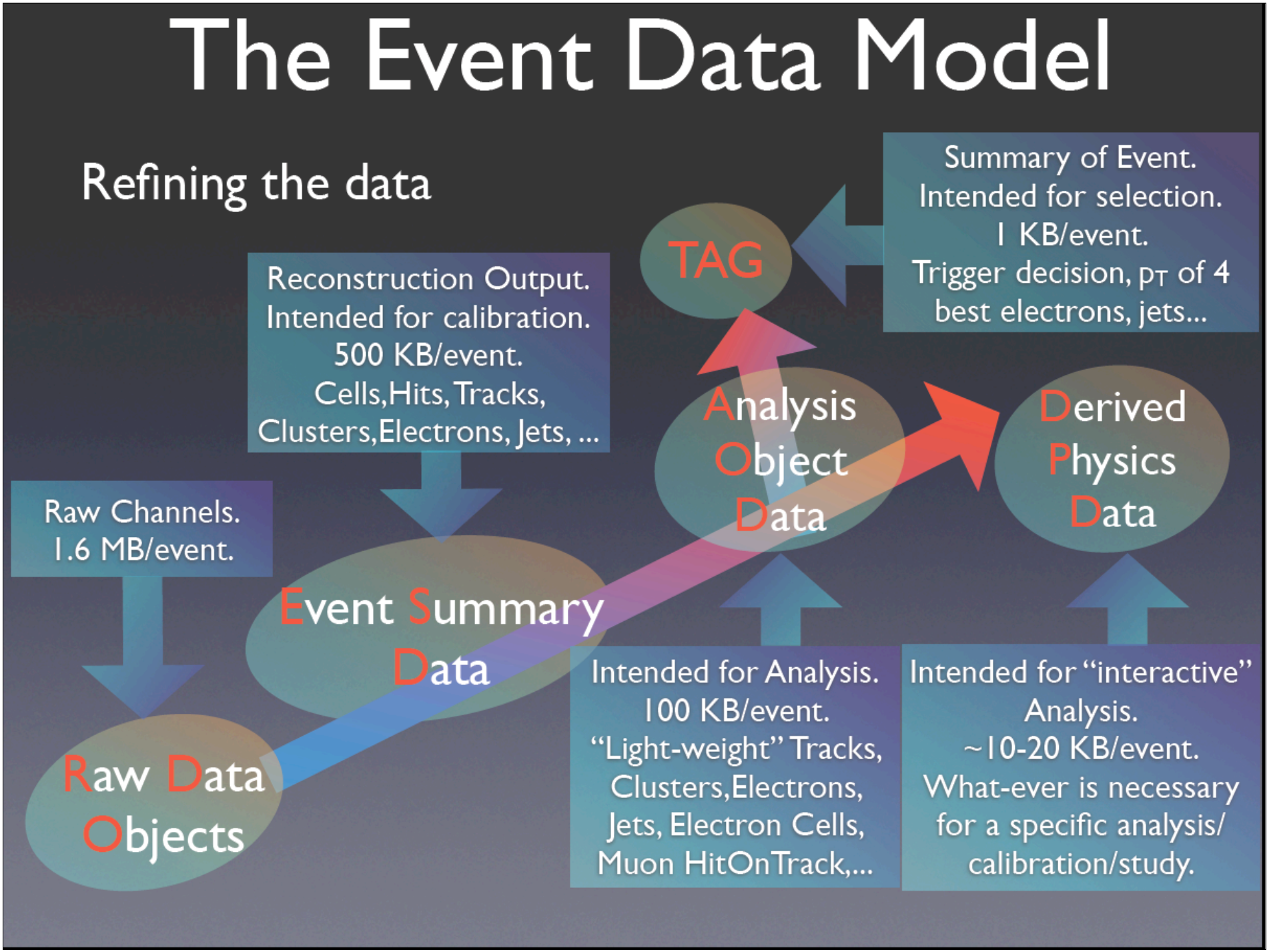
MHz
Interactive Analysis

Statistical Analysis

Data Analysis & Calibration

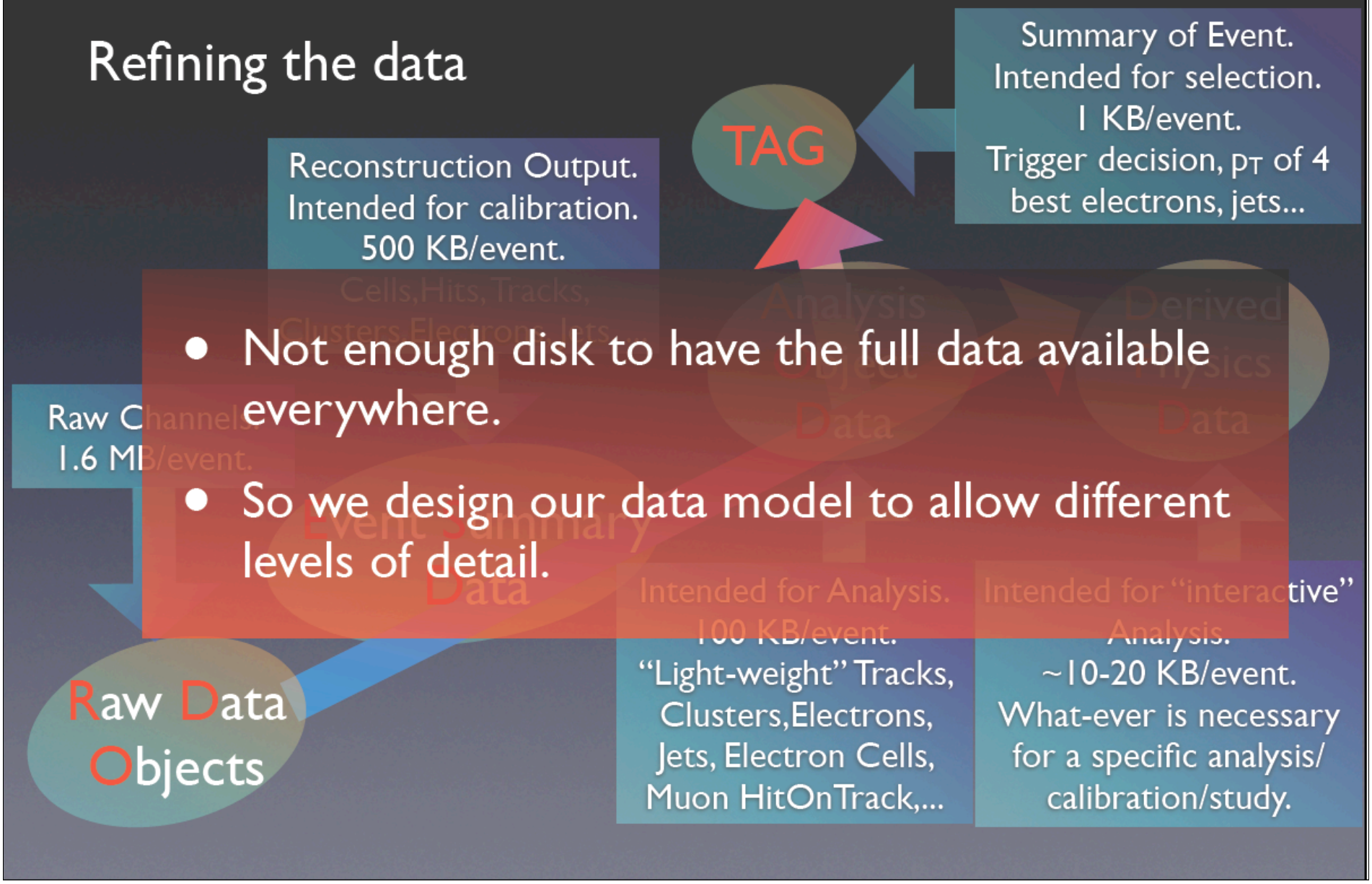
The Event Data Model

Refining the data



The Event Data Model

Refining the data



- Not enough disk to have the full data available everywhere.
- So we design our data model to allow different levels of detail.

The Computing Model

• Resources Spread Around the GRID

•Interactive Analysis
•Plots, Fits, Toy MC, Studies, ...

Tier 3

DPD

•Reprocessing of full data with improved calibrations 2 months after data taking.
•Managed Tape Access: RAW, ESD
•Disk Access: AOD, fraction of ESD

30 Sites Worldwide

Tier 2

AOD

•Derive 1st pass calibrations within 24 hours.
•Reconstruct rest of the data keeping up with data taking.

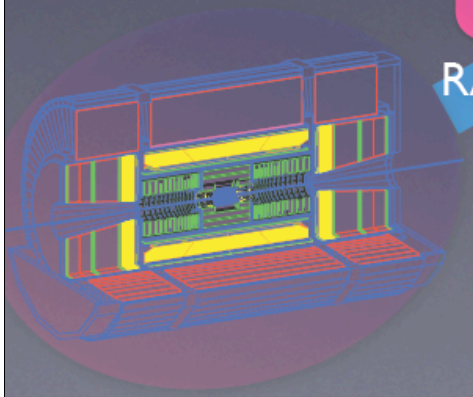
Tier 1

10 Sites Worldwide

RAW/ AOD/ ESD

Tier 0

RAW

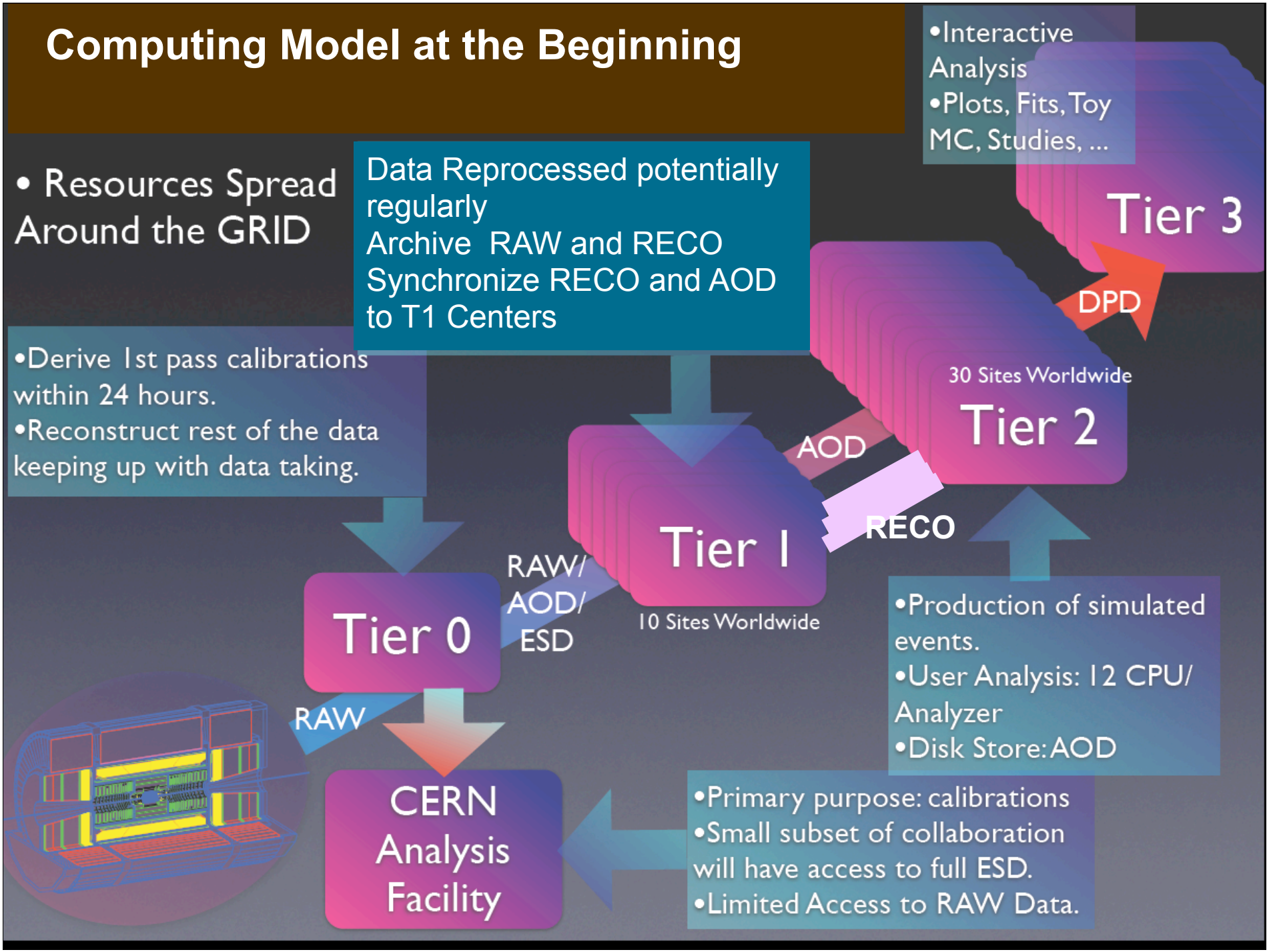


CERN Analysis Facility

•Production of simulated events.
•User Analysis: 12 CPU/ Analyzer
•Disk Store: AOD

•Primary purpose: calibrations
•Small subset of collaboration will have access to full ESD.
•Limited Access to RAW Data.

Computing Model at the Beginning



Analysis Activity

- *Re-reconstruction/re-calibration*- CPU intensive... often necessary.
- *Algorithmic Analysis*: Data Manipulations
  ESD→AOD→DPD→DPD
  - *Skimming*- Keep interesting events
  - *Thinning*- Keep interesting objects in events
  - *Slimming*- Keep interesting info in objects
  - *Reduction*- Build higher-level data which encapsulates results of algorithms
  - Basic principle: Data Optimization + CPU intensive algs → more portable input & less CPU in later stages.

- *Tier 1/2 Activity*
  - Framework (ie Athena) based
  - Resource intensive
  - Large scale (lots of data)
  - Organized
  - ***Batch access only***

Primary difference

- *Interactive Analysis*: Making plots/performing studies on highly reduced data.
- *Statistical Analysis*: Perform fits, produce toy Monte Carlos, calculate significance.

- *Tier 3 Activity*
  - Often exo-framework
  - ***Interactive***

Network Estimates

- From the CMS Computing Model (ATLAS is slightly higher):
- The network requirements for Tier-0 to Tier-1 transfers are driven by the trigger rate and the event size
  - Estimates are ~2.5Gb for a nominal Tier-1 center
    - The Tier-1 event share with a factor of 2 recovery factor and a factor of 2 provisioning factor
- The Tier-1 to Tier-1 transfers are driven by the desire to synchronize re-reconstruction samples within a short period of time
  - To replicate the newly created reconstructed and AOD between Tier-1 centers in two week is 1Gb/s)
- The Tier-1 to Tier-2 transfers are less predictable
  - Driven by user activities.
  - CMS model estimates this at 50-500MB/s (Includes safety factors)

Tier-1 to Tier-2 Connectivity

- In order to satisfy their mission as a primary resource for experiment analysis the Tier-2 need good connectivity to the Tier-1 centers
  - Data is served from Tier-1 computing centers.
    - In CMS Each Tier-1 is assigned a share
    - In ATLAS the Tier-1s have a complete analysis
- The connectivity between the Tier-1 and Tier-2 centers can be substantially higher than the Tier-0 to Tier-1 rates
  - Already in the computing challenge the incoming rate to FNAL is half the outgoing rate to Tier-2 centers
  - The network that carries the Tier-2 traffic is going to be instrumental to the experiment's success.
- The Tier-2 traffic is a more difficult networking problem
  - The number of connections is large
  - There are a diverse set of locations and setups

Tier-2 and Tier-3 Centers

- A Tier-2 center in ATLAS and CMS are approximately 1MSI2k of computing
  - Tier-3 centers belong to university groups and can be of comparable size
- A Tier-2 center in ATLAS and CMS ~200TB of disk
  - Currently procuring and managing this volume of storage is expensive and operationally challenging
    - Requires a reasonably virtualization layer
- A Tier-2 center has between 2.5 Gb/s and 10Gb/s of connectivity in the US
  - This is similar between Tier-2 and Tier-3 centers
  - The speed of connection to the local sites has increased rapidly
- In the US-CMS planning a Tier-2 supports 40 Physicists performing

Surviving the first years

- The computing for either experiment is hardest as the detector is being understood
  - The AOD for both experiments is 100kB
    - An entire year's data and simulation is only ~300TB
  - Data is divided into ~10-20 trigger streams
  - A physics analysis should rely on 1 trigger stream
    - A Tier-2 could potentially maintain all the analysis objects for the majority of the analysis streams
- Unfortunately, until the detector and reconstruction are completely understood the AOD is not useful for most analysis and access to the raw data will be more frequent
  - The full raw data is 15-20 times bigger
  - Transition will be smooth, but may take years to complete
  - People working at Tier2 centers can make substantial, but bursty requirements of the data transfers

First Year and Schedule

- In 2008 we expect colliding beams in the late summer
  - The run will be at a lower energy than the design, but still a new energy frontier
    - 10TeV at the beginning
- The run is likely to be reasonably short
  - Current estimate is around 45 days
  - Collect 40-50pb-1
  - Accelerator live time is probably between 10-20%
  - Somewhere between 200-300M events
- The experiments will write out as much as they can
  - A lot of the events may be less interesting in the future

First Year of Data

- Rough calculation of data volume
- Assuming the data rate is driven by live time and not luminosity
  - Use 45 day run
  - Total Lumi 40-50pb-1
    - **Luminosity numbers maybe optimistic**
  - Assumes roughly 20% livetime
  - Assume 20% overlap in primary dataset definitions, which leads to 20% more data
  - Assumes 0.5MB event for ESD/RECO
  - 100kB events for AOD
  - Assume 300Hz
  - Assume MC equal size to data
- Works out to roughly 200TB of DATA and 200TB of MC

Analysis Selections

- When going back to the RECO data and complete simulation, analysis selections on a complete trigger streams
  - 1% selection on data and MC would be 0.4TB, 10% selection would be 4TB
    - Assumes all Primary Datasets are the sample size
  - There are an estimated 40 people working at a Tier-2
    - If half the people perform the small selections at the level of twice a month
      - **This is already 50MB/s on average and everyone is working asynchronously**
      - **The original analysis estimates were once a week**
      - **Some primary datasets will be larger**
    - 100MB/s x 7 Tier-2s would be ~6Gb/s from a Tier-1
- Size of selections, number of active people and frequency of selections all have significant impact on the total network requirements
  - Can easily arrive at 500MB/s for bursts.
  - High end for computing model numbers

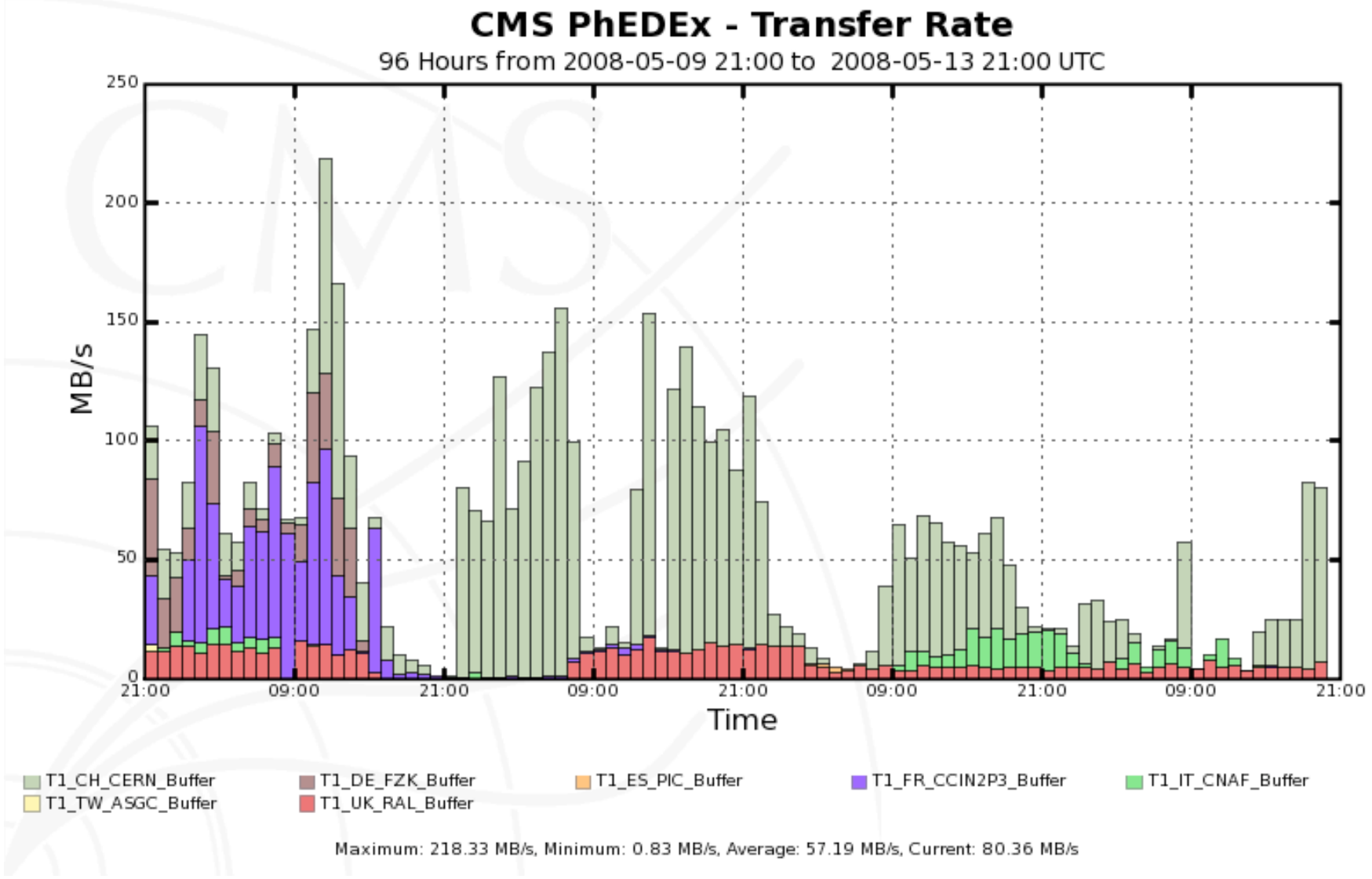
Top Down Network Estimate

- When US-CMS made the initial network requirements of the Tier-2 centers they were made with two scales in mind
  - A user would be working with multi-terabyte samples and they would need to be moved within a day.
    - We now have networks that can do this. We're still working on computing infrastructure to do it reliably with low latency
  - A center would have 200TB of disk
    - The experiment, especially at the beginning, would be re-reconstructing regularly and causing data to be updated
    - Flushing 200TB of disk
      - **100Mb/s is 200 days (essentially static)**
      - **1Gb/s is 20 days**
      - **10Gb/s is 2 days**
    - In the end more than 1Gb/s seems acceptable but preferably closer to 10Gb/s

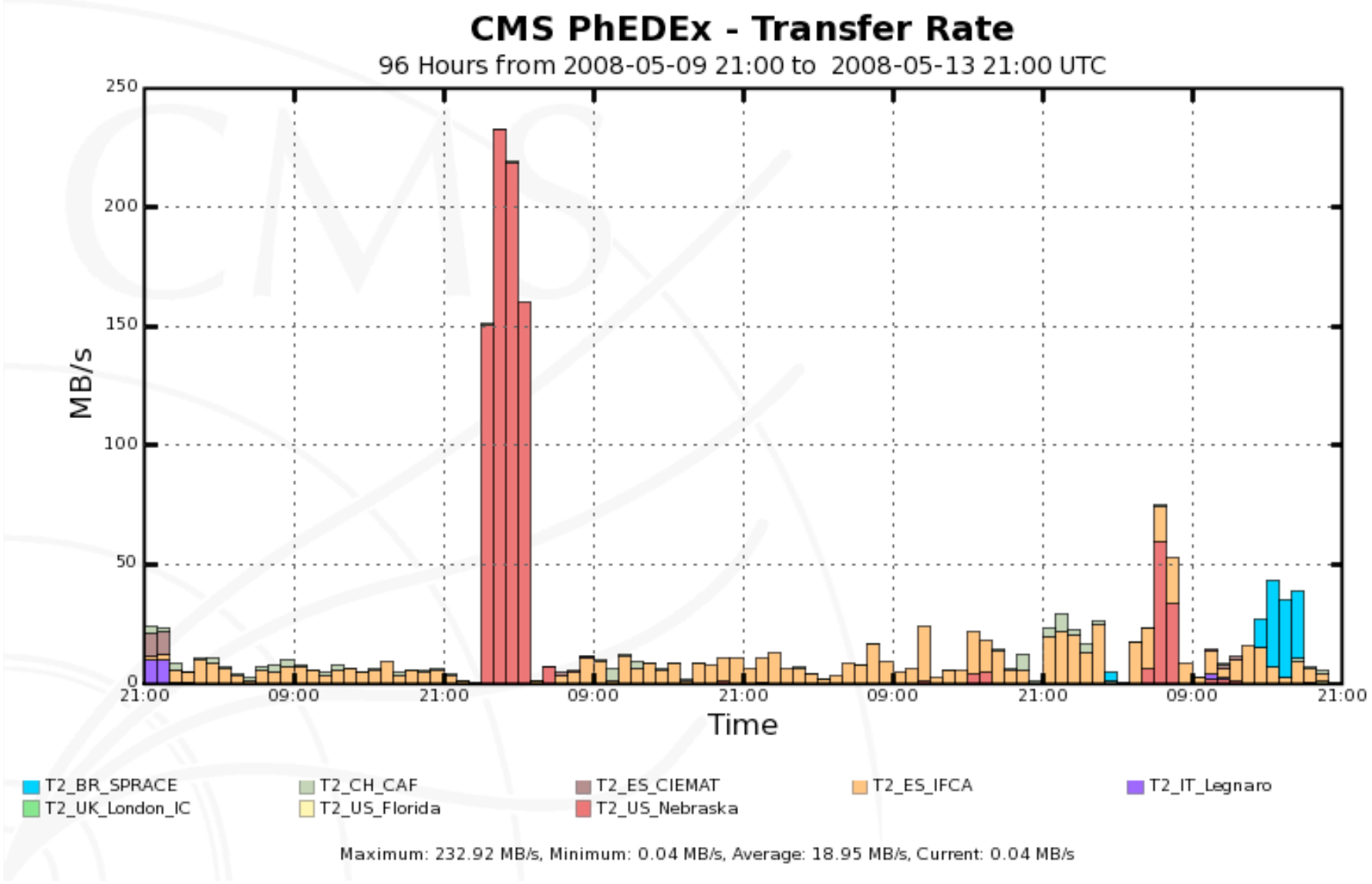
Tier-3 Connectivity

- Tier-2s are a resource for the physics community.
  - Even people with significant university clusters at home have the opportunities to use Tier-2
  - The use of Tier-3s for analysis is foreseen in the model
    - These are not resources for the whole experiment and can have lower priority for access to common resources
- The number of active physicist supported at a Tier-3 center is potentially much smaller than a Tier-2
  - 4-8 people
  - This leads to smaller sustained network use
    - but similar requirements to T2s to enable similar turn-around times/ latencies for physics datasets copied to T3 sites for analysis
- For CMS, the Tier-3s are similar in analysis functionality but not in capacity

Moving Data Around to T1



Moving Data Around T2



Outlook

- We expect that this year the LHC will take data
  - It is unlikely to be a long running period but it will collect very interesting data at an energy never produced at an accelerator
- The potentially smaller raw event sample will not decrease the network utilization by either experiment
  - At the beginning the data will be frequently reprocessed as the detectors are being commissioned
  - At the beginning the user communities may need to look at larger event formats, these can easily counteract the initially smaller number of events
- The Tier-2 and Tier-3 centers will be the most important resource for experiment analysis
  - They will require good connectivity to complete that mission